Prospects for cost-effective genomic selection via accurate within-family imputation

Published December 22, 2016

RESEARCH

Prospects for Cost-Effective Genomic Selection via Accurate Within-Family Imputation

Gregor Gorjanc,* Mara Battagin, Jean-Francois Dumasy, Roberto Antolin, R. Chris Gaynor, and John M. Hickey

ABSTRACT

Genomic selection has great potential to increase the efficiency of plant breeding, but its implementation is hindered by the high costs of collecting the necessary data. In this study we evaluated the potential of accurate within-family imputation for enabling cost-effective genomic selection. We have simulated a breeding program with inbred parents and their segregating progeny distributed among families, of which some were used as a training set and some were used as a prediction set. Parents were genotyped at high density (20,000 markers), while progeny were genotyped at high or low density (500, 200, 100, or 50 markers) and imputed. Low-density markers were chosen to segregate within each family separately. The assumed low-density genotyping costs accounted for this assumption. Six sets of scenarios were analyzed in which imputation was leveraged to maximize cost effectiveness of genomic selection by (i) decreasing the genotyping costs, (ii) increasing selection intensity by genotyping more individuals at fewer markers, or (iii) increasing prediction accuracy by genotyping more phenotyped individuals at fewer markers. The results show that, with a constant size of the training and prediction sets, the prediction accuracy was unimpaired when at least 200 low-density markers were used. However, the return on investment was maximal (5.67 times that of the baseline scenario) when only 50 low-density markers were used because that enabled maximal reduction in the genotyping costs and only minimal reduction in the prediction accuracy. Increasing either the training set or prediction set further increased the return on investment when imputed genotypes were used, but not when the true high-density genotypes were used. The results show how plant breeding programs can implement genomic selection in a cost-effective way.

G. Gorjanc, M. Battagin, J.-F. Dumasy, R. Antolin, R.C. Gaynor, and J.M. Hickey, The Roslin Institute and Royal (Dick) School of Veterinary Studies, Univ. of Edinburgh, Easter Bush Research Centre, Midlothian EH25 9RG, UK. Received 16 June 2016. Accepted 9 Nov. 2016. *Corresponding author (Gregor.Gorjanc@roslin.ed.ac.uk). Assigned to Associate Editor Aaron Lorenz.

This study evaluates the potential of accurate within-family imputation for enabling cost-effective genomic selection in plant breeding. Genomic selection has great potential to increase the efficiency of plant breeding (Bernardo and Yu, 2007; Lorenzana and Bernardo, 2009). Perhaps most importantly, genomic selection increases the accuracy of early assessment of the genetic merit and therefore enables rapid recurrent selection. In practice, implementing genomic selection can be challenging due to the high costs of collecting the necessary amounts of data, which must meet a set of requirements and must integrate with the breeding program.

Two large sets of data are required for the full exploitation of the potential of genomic selection; a training set of genotyped and phenotyped individuals and a prediction set of genotyped-only individuals (Meuwissen et al., 2001; Daetwyler et al., 2008; Goddard, 2009). The training set is used to estimate parameters of the genomic selection model. To accurately estimate the parameters, the training set must be comprised of a large number of genotyped and phenotyped individuals. The prediction set represents the selection candidates, whose genetic merit will be predicted. Ideally, the prediction set would be large, because this enables high selection intensity and consequently high response to selection. However, large training and prediction sets increase costs that must be balanced against the potential increase in response to selection. The two sets of data can be assembled in different ways, and this has important implications for use of genomic selection in plant breeding.

Published in Crop Sci. 57:1–13 (2017).
doi: 10.2135/cropsci2016.06.0526

Initial proposals of genomic selection in plant breeding suggested training and prediction within a family (Bernardo and Yu, 2007; Lorenzana and Bernardo, 2009). Such an implementation requires small amounts of data but does not fully utilize the potential of genomic selection. For example, to achieve genomic prediction accuracy of 0.5 in a biparental family, a training set should be composed of about 100 phenotyped individuals from the family that are genotyped at a few hundred markers (Bernardo and Yu, 2007; Lorenzana and Bernardo, 2009; Hickey et al., 2014; Lian et al., 2014). These low requirements are due to the limited genetic diversity within a family and high relatedness between the training and prediction individuals within a family (Daetwyler et al., 2008; Goddard, 2009; Pszczola et al., 2012; Hickey et al., 2014). However, assembling the training set within a family is time consuming and delays the use of genomic selection until the late generations of genetic improvement within a family. The potential for genomic selection at that stage is lessened in comparison with early stages and phenotypic selection (e.g., Endelman et al., 2014; Jacobson et al., 2014).

Recent proposals of genomic selection in plant breeding suggested training and prediction across families (Heffner et al., 2011; Hickey et al., 2014; Jacobson et al., 2014; Mackay et al., 2014). Such an implementation requires large amounts of data and fully utilizes the potential of genomic selection. For example, to achieve genomic prediction accuracy of 0.5 in a new biparental family, a training set should be composed of at least a few thousand phenotyped individuals from other families that are genotyped with about 10,000 markers (Hickey et al., 2014). These requirements are due to more diversity among families than within a family and potentially low relatedness between the training and prediction individuals (Daetwyler et al., 2008; Goddard, 2009; Clark et al., 2012; Pszczola et al., 2012; Hickey et al., 2014).

While training across families with a large number of densely genotyped individuals can be significantly more expensive than training within a family with a small number of sparsely genotyped individuals, it provides important advantages for plant breeding programs. Most importantly it enables selection for quantitative traits, such as yield, in early generations of segregating populations. This enables a reduction in generation interval and an increase in selection intensity, which are the key advantages of genomic selection in comparison with phenotypic selection (e.g., Schaeffer, 2006; Bernardo and Yu, 2007; Gaynor et al., unpublished data, 2016). Additionally, training the genomic selection model across families enables continuous expansion and updating of the training set with data from each new family. Such an expansion increases the prediction accuracy and reduces its sampling variance (Hickey et al., 2014), which reduces the variance of response to selection (Nicholas, 1980). In addition, reuse of the collected data increases its value and distributes the costs of setting up the training set over a longer time period and a larger number of predictions.

Several studies have suggested leveraging the power of imputation for genomic selection in plant breeding (e.g., Hickey et al., 2012a; Rutkoski et al., 2013; He et al., 2015; Xavier et al., 2016). However, most of these studies used relatively inaccurate imputation methods and genotyping strategies that do not explicitly leverage the family structure of plant breeding programs, such as those used by Hickey et al. (2015). In addition, the studies did not quantify the potential of imputation to reduce the cost of assembling the required data for genomic selection (Huang et al., 2012; Cleveland and Hickey, 2013; Jacobson et al., 2015). Plant breeders could leverage imputation in several ways to maximize the return on investment in genomic selection. First, breeders could reduce the cost of each prediction by genotyping selection candidates at a few markers and imputing the untyped markers. Perhaps the same approach could also be used to reduce the cost of assembling the training set. Second, given a fixed genotyping budget, breeders could increase response to selection by genotyping more selection candidates with fewer markers and imputing the untyped markers with the aim to increase selection intensity. Third, breeders could also increase response to selection by genotyping more training individuals with fewer markers and imputing the untyped markers with the aim to increase the genomic prediction accuracy. Both the second and the third option would utilize existing phenotypes at a fraction of full genotyping costs and would therefore increase the return on investment in both phenotype and genotype data. Finally, the enlarged training and prediction sets could be used jointly with the aim to increase response to selection via both increased selection intensity and accuracy.

The aim of this study was to evaluate the potential of accurate within-family imputation for enabling cost-effective genomic selection in plant breeding. We addressed this by evaluating:

(i) the prediction accuracy and return on investment with imputed genotypes in the prediction set and/or the training set

(ii) the response to selection and return on investment with imputed genotypes in an enlarged prediction set

(iii) the prediction accuracy and return on investment with imputed genotypes in an enlarged training set

(iv) the response to selection and return on investment with imputed genotypes in enlarged training and prediction sets.

MATERIALS AND METHODS

We used stochastic simulation to evaluate the potential of accurate within-family imputation to enable cost-effective genomic selection in plant breeding. The simulation involved the following steps, most of which were performed with the

AlphaSim program (Faux et al., 2016), available at http://www.AlphaGenes.Roslin.ed.ac.uk/AlphaSuite/AlphaSim:

(i) generating founder genomes

(ii) selecting causal loci, defining trait architecture, and selecting several marker arrays

(iii) generating a breeding program with inbred parents and a series of segregating families

(iv) applying a cost-effective genotyping strategy (densely genotype the parents and sparsely genotype their progeny) and within-family imputation

(v) training the genomic selection model and predicting breeding values in a range of scenarios

(vi) describing results within each scenario with prediction accuracy or response to selection and return on investment.

Obtained results were summarized over 30 replicates and presented graphically, while Supplemental Table S2 provides results in the tabular form. Data preparation and summaries were performed with the R program (R Development Core Team, 2014).

Genome

The genome was simulated by sampling 100 haplotype sequences for each of 10 chromosomes using the Markovian Coalescent Simulator (MaCS) (Chen et al., 2009). Each chromosome was 100 cM long and included 1×10^8 base pairs. Chromosomes were simulated using a per-site recombination rate of 1×10^{-8}, a per-site mutation rate of 1×10^{-8}, and an effective population size that varied over time. The effective population size was set to 50 in the final generation of the coalescent simulation, to 100 at 10 generations ago, to 1000 at 100 generations ago, to 6000 at 1000 generations ago, to 12,000 at 10,000 generations ago, and to 32,000 at 100,000 generations ago, with linear changes in between. The resulting genome sequences had approximately 1000,000 segregating variants (bi-allelic single-nucleotide polymorphisms) in total.

Causal Loci, Phenotypes, and Marker Arrays

A quantitative trait was simulated as being influenced by 10,000 loci sampled at random from the segregating sequence variants with the restriction of an equal number from each chromosome. These causal loci had additive effects sampled from a normal distribution with a mean of zero and variance of one divided by the number of loci. True breeding value of an individual was calculated as the sum of additive effects of alleles at the causal loci that the individual inherited. Phenotype value of an individual was sampled from a normal distribution with a mean equal to the true breeding value of the individual and a residual variance according to the heritability. The heritability and the residual variance were computed relative to the additive genetic variance in the base population.

One high-density and four low-density marker arrays were constructed. The high-density array had 20,000 markers in total (2000 markers per chromosome) sampled from the non-causal segregating sequence variants with the restriction of an equal number from each chromosome. The low-density arrays had 50, 100, 200, or 500 markers in total with 5, 10, 20, or 50 markers per chromosome, respectively. Markers for the low-density arrays were selected at random from the high-density array with two restrictions. First, they were specific for each family by selecting markers with opposing homozygous genotypes in parents of a family. We did this to get an exact number of segregating markers within each family. This design choice was accounted for when we evaluated the genotyping costs. We refer to the number of segregating markers throughout the manuscript unless otherwise stated. Second, the markers were nested, i.e., markers on the smallest low-density array were present on the second smallest low-density array, etc.

Breeding Program

A breeding program of a self-pollinating species with inbred parents and biparental populations (families) was simulated (Fig. 1). The program was initiated by establishing a base population of 40 inbred parents. Each parent had one haplotype per chromosome sampled from the base haplotypes, allowing for recombination between base haplotypes. The sampled haplotypes were doubled to create inbreds. The parents were then crossed at random to create 160 biparental populations, with a restriction that the same parents could only be crossed once. Each biparental population was developed by selfing the F_1 individual to generate 200 F_2 individuals, who were further selfed to generate 200 F_3 individuals that gave rise to 200 $F_{3:4}$ populations. These were evaluated in a preliminary yield trial with a heritability of 0.1. This phenotype pertained to the

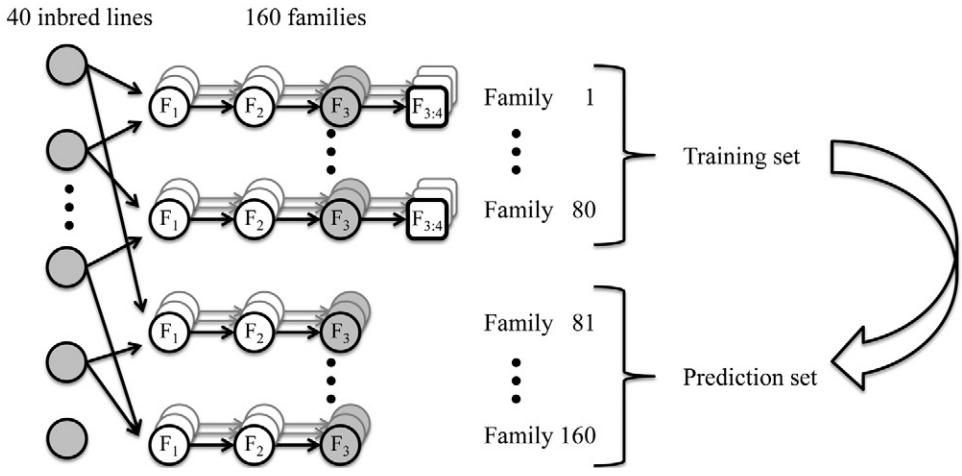

Fig. 1. Breeding program design with 40 inbred lines used to generate 160 families, of which 80 families comprised a genomic selection training set and 80 families comprised a prediction set. Circles represent individuals (shaded had genotype data) and squares represent phenotypic data

genotype of the F_3 individual used to derive the $F_{3:4}$ population. In practice, the development of a biparental population would continue for a number of further generations, but for this study, it was stopped at this stage, as all the individuals and phenotypes required for genomic selection, according to the design that we chose to use, had been generated.

Genotyping Strategy and Imputation

All parents were genotyped with the high-density array, while F_3 individuals were either genotyped with the high-density array or genotyped with one of the low-density arrays and imputed to the high-density array. We performed imputation with a slightly modified version of the method of Li et al. (2010), which was implemented in a new version of the AlphaImpute program (Hickey et al., 2012b, available at http://www.AlphaGenes.Roslin.ed.ac.uk/AlphaSuite/AlphaImpute). The imputation method involved (i) constructing a set of template haplotypes, (ii) estimating model parameters that describe the mapping of the observed genotypes onto the template haplotypes, and (iii) estimating (imputing) genotype probabilities and allele dosages at untyped markers for sparsely genotyped individuals. We used estimated allele dosages as imputed genotypes. Each family and chromosome was imputed independently. Inputs for imputation were the high-density genotypes of the two crossing parents and low-density genotypes of the 200 F_3 individuals. Accuracy of imputation was measured with the Pearson correlation between the standardized imputed allele dosages and the standardized true genotype at untyped markers; the correlation was computed one individual at a time and averaged over individuals (Hickey et al., 2012a; Calus et al., 2014). Standardization included centering (subtracting average allele dosage, which is equal to two times the allele frequency) and scaling (dividing by the standard deviation of allele dosages, which is equal to square root of heterozygosity). We also report the unstandardized accuracies in parentheses for completeness. Preliminary tests showed limited imputation accuracy with the method of Li et al. (2010): 0.09 (0.58) with 5 markers per chromosome, 0.31 (0.71) with 10 markers per chromosome, 0.60 (0.83) with 20 markers per chromosome, and 0.77 (0.89) with 50 markers per chromosome. This is expected, as the high-density genotype for each family was only available on two parents, which provides limited information to construct an informative set of template haplotypes and estimate model parameters (Li et al., 2010).

We improved the initially low imputation accuracy by leveraging the inbred status of parents; *in silico*, we generated 98 doubled haploid F_2 individuals from the parental genotypes (assuming uniform recombination rates, although in practice, any recombination map could be used) and used them to expand the set of template haplotypes and improve estimates of model parameters. This procedure improved accuracy of imputation to 0.61 (0.81) with 5 markers per chromosome, 0.76 (0.89) with 10 markers per chromosome, 0.86 (0.93) with 20 markers per chromosome, and 0.93 (0.96) with 50 markers per chromosome. Between families, there were some differences in the imputation accuracy. The range was 0.56 to 0.67 (0.75 to 0.88) with 5 markers per chromosome, 0.66 to 0.81 (0.81 to 0.95) with 10 markers per chromosome, 0.82 to 0.88 (0.87 to 0.97) with 20 markers per chromosome, and 0.88 to 0.95 (0.91 to 0.98) with 50 markers per chromosome.

The computational time to impute 2000 markers per chromosome for 200 individuals was about 10 min with 200 template haplotypes and 100 iterations. Preliminary analyses showed that increasing number of low-density markers, number of template haplotypes, and iterations increased accuracy. We also observed an interaction between the number of template haplotypes and iterations, i.e., accuracy can be increased by iterating over a few template haplotypes many times or iterating over many template haplotypes fewer times. Generally, 10 to 20 iterations gave high imputation accuracy that was only marginally improved in further iterations. On the other hand, computational time for certain level of accuracy increased with reduced number of low-density markers and increased number of template haplotypes and iterations. Analysis of the implemented algorithm shows that the computational time is quadratic in the number of haplotypes and linear in the number of iterations. The chosen setting gave accurate imputations with acceptable computational time.

Genomic Prediction

Genomic predictions of breeding values for genotyped-only F_3 individuals within a family were based on estimated marker associations from training on other families (Fig. 1). The size and composition of the training set varied between and within the different scenarios. Marker associations were estimated by regressing phenotypic values on allele dosages with the ridge regression model (Hoerl and Kennard, 1976; Whittaker et al., 2000; Meuwissen et al., 2001) as implemented in the AlphaBayes program, available at http://www.AlphaGenes.Roslin.ed.ac.uk/AlphaSuite/AlphaBayes. The model parameters were estimated using a Monte Carlo Markov Chain method with one chain of 10,000 iterations, of which the first 1000 were discarded as burn-in. Posterior means were used as estimates of marker associations.

Prediction Accuracy

Accuracy of genomic prediction was measured with the Pearson correlation between predicted and true breeding values. We measured accuracy in two ways, jointly across families and within each family, to remove the between-family source of variation (e.g., Windhausen et al., 2012). In the results, we refer to this as the scope of prediction. The within-family correlation measures accuracy of predicting the within-family variation, commonly referred to as Mendelian sampling variation. The across-family correlation measures accuracy of predicting the within-family and between-family variation. The aim of genomic prediction is to capture variation due to both components, but it is harder and more important to capture the within-family variation, as it is this component that drives sustainable genetic gain (e.g., Woolliams et al., 1999; Hickey et al., 2014). We therefore focus largely on the accuracy within a family in the results and discussion but report both for completeness.

Response to Selection

Response to selection was measured only for selection within a family for the same reasons as described for accuracy (see previous paragraph). It was calculated by subtracting the mean true breeding value of selection candidates within a family from the mean true breeding value of the 10 selected individuals. Selection was based on genomic predictions of breeding values.

Return on Investment

Return on investment was measured by dividing the response to selection within a family by the accrued genotyping costs to

achieve that response to selection. We expressed it relative to a chosen baseline so that all the evaluated scenarios could be compared. We considered only the costs of genotypes, as we assumed that a breeding program would already have phenotypes available. For simplicity, other costs were ignored. We divided the cost of training genotypes by 80, because we performed predictions in 80 families and all of them used the same training data. We believe this is a conservative choice, as a real breeding program could spread this cost over many more families generated in several cycles of genomic selection. The cost of prediction genotypes was considered for each family separately, because the response to selection was measured for each family separately.

We assumed that the cost of a high-density array with 20,000 markers is US$30.00. Further, we assumed that the cost of a low-density array is due to fixed and variable components. The fixed component was set to $2.50, while the variable component was set to $1.00 for 100 markers. Because the low-density markers were chosen for each family, we assumed that the total number of markers on a low-density array would have to be three times larger (e.g., Hickey et al., 2014), and we factored this into the costs. The cost of low-density arrays therefore ranged between 13 and 58% of the high-density array (Table 1). We provide a spreadsheet in the supplement that details the calculations (Supplemental Table S2), which can be used to modify our cost assumptions for genotypes and phenotypes.

Scenarios

We analyzed the simulated data in six sets of scenarios in which imputation was leveraged to maximize utility of genomic data in our chosen breeding program design (Table 2, Supplemental Table S2). Across the scenarios, the utility of the resources was maximized based on three principles: (i) decreasing the genotyping costs, (ii) trading off selection intensity versus prediction accuracy by genotyping more or fewer individuals at fewer or more markers, and (iii) increasing prediction accuracy by genotyping more or fewer phenotyped individuals at fewer or more markers, i.e., enlarging the training set at the expense of the precision in the genotyping of each individual in the training set. Each scenario involved (i) constructing a training set with the true or imputed genotypes comprised of individuals from a number of families, (ii) estimating parameters of the genomic selection model, and (iii) predicting breeding values with the true or imputed genotypes in a prediction set of a distinct set of families.

Table 1. Assumed costs of high-density and low-density genotype data.

Number of markers	Cost	Ratio
	US$	
High-density		
20,000	30.00	1.00
Low-density		
500	17.50	0.58
200	8.50	0.28
100	5.50	0.18
50	4.00	0.13

Table 2. Summary of scenarios.

Scenario	Training set†	Prediction set	Result‡
1	HD	LD	Accuracy & ROI
2	LD	HD	Accuracy & ROI
3	LD	LD	Accuracy & ROI
4	HD	LD & enlarge	Response to selection & ROI
5	LD & enlarge	HD	Accuracy & ROI
6	LD & enlarge	LD & enlarge	Response to selection & ROI

† HD, high-density genotypes; LD, low-density genotypes imputed to high density.

‡ ROI, return on investment.

The first three scenarios quantified the prediction accuracy and return on investment by using imputation with different numbers of low-density markers in the training and/or prediction set (Table 2, Supplemental Table S2). The first scenario used high-density genotypes in the training set and low-density genotypes imputed to high density in the prediction set. The second scenario used low-density genotypes imputed to high density in the training set and high-density genotypes in the prediction set. The third scenario used low-density genotypes imputed to high density in both the training and prediction set. The training set comprised 80 families, with each family contributing 25 training individuals. In total, this gave a training set with $80 \times 25 = 2000$ individuals (Supplemental Table S2). The families and individuals within families were selected at random among all the available families and individuals within families. Predictions were performed in the other 80 families, for 200 F_3 individuals within each family.

The fourth scenario quantified the response to selection and return on investment in an enlarged prediction set genotyped at fewer markers (Table 2, Supplemental Table S2). Since we always selected the fixed number of individuals, the change in the size of the prediction set (the number of selection candidates) translates to the increased selection intensity. The following four strategies that had approximately the same cost were evaluated (denoted as x selection candidates genotyped at y low-density markers—xI@yM): 50I@500M, 100I@200M, 150I@100M, and 200I@50M (Supplemental Table S2). When a strategy did not involve genotyping all of the potential selection candidates of a family, a random sample of candidates was taken. The training set was the same as in the first scenario and was genotyped at high density. The prediction set had either the true high-density genotypes or low-density genotypes imputed to high density.

The fifth scenario quantified the prediction accuracy and return on investment in an enlarged training set genotyped at fewer markers (Table 2, Supplemental Table S2). The same four strategies were used as in the fourth scenario but were applied to the training set. The following four strategies that had approximately the same cost were evaluated (denoted as 80 families times x individuals per family genotyped at y low-density markers—80C×xI@yM): 80C×50I@500M, 80C×100I@200M, 80C×150I@100M, and 80C×200I@50M. When a strategy did not involve genotyping all of the individuals within a family, a random sample of individuals was taken. In total, the training set had 4000 individuals for the strategy 80C×50I@500M, 8000 individuals for the strategy 80C×100I@200M, 12,000 individuals for the strategy 80C×150I@100M, and 16,000 individuals for the strategy 80C×200I@50M (Supplemental Table S2). The training set therefore had low-density genotypes imputed to high density. The true high-density genotypes were also used for comparison.

The prediction set was the same as in the first scenario, but this time with the true high-density genotypes.

The sixth scenario quantified the response to selection and return on investment in an enlarged training and prediction set genotyped at fewer markers (Table 2, Supplemental Table S2). This scenario was a combination of the fourth and the fifth scenarios with exactly the same setting, with the only difference that low-density genotypes imputed to high density were used both in training and prediction. The true high-density genotypes were also used for comparison.

RESULTS

This paper uses simulation to evaluate the prospect of accurate imputation to enable cost-effective genomic selection in plant breeding. The results show that accurate imputation can enable cost-effective genomic selection through (i) reduction of genotyping costs in training and prediction sets, (ii) increase of selection intensity by enlarging the prediction set, and (iii) increase of prediction accuracy by enlarging the training set. These advantages enable breeders to increase the return on investment in the required data for genomic selection.

Prediction Accuracy with Imputation in Training and/or Prediction

Prediction accuracy decreased marginally with the decreasing number of low-density genotypes used for imputation. This is shown in Fig. 2, which plots the genomic prediction accuracy against the number of markers used in the prediction set. The training set had high-density genotypes. Accuracies are shown both for prediction across families and within a family. The baseline accuracy with the true high-density genotypes was 0.71 for prediction across families and 0.42 for prediction within a family. The accuracy decreased with the decreasing number of low-density genotypes. The decrease was less pronounced for prediction across families (from 0.71 to 0.66) than for within a family (from 0.42 to 0.34). Similar trends were observed when imputation was used in the training set and not in the prediction set (scenario 2; Supplemental Table S1) and when imputation was used in both sets (scenario 3; Supplemental Table S1). Supplemental Table S1 also shows that the main difference between the three scenarios was in the rate of the decrease in accuracy. The rate was lowest when imputation was used only in the training set (scenario 2).

Using imputed genotypes gave greater return on investment than using the true high-density genotypes. This is shown in Fig. 3, which plots the return on investment of selecting within a family against the number of markers used in the training and prediction set in the first three scenarios. The baseline for comparison was a strategy where both the training and prediction set had high-density genotypes. Return on investment increased with the decreasing number of low-density markers and there were large differences between the scenarios. The greatest increases were observed when imputed genotypes were used both in the training and prediction set. In that scenario, the greatest return on investment was 5.67 times that of the baseline scenario when we used only 50 low-density markers. Intermediate increases were observed when imputed genotypes were used only in the prediction set. In that scenario, the greatest return on investment was 3.52 times that of the baseline scenario when we used only 50 low-density markers. The lowest increases were observed when imputed genotypes were used only in the training set. In that scenario, there were no significant differences between the marker densities.

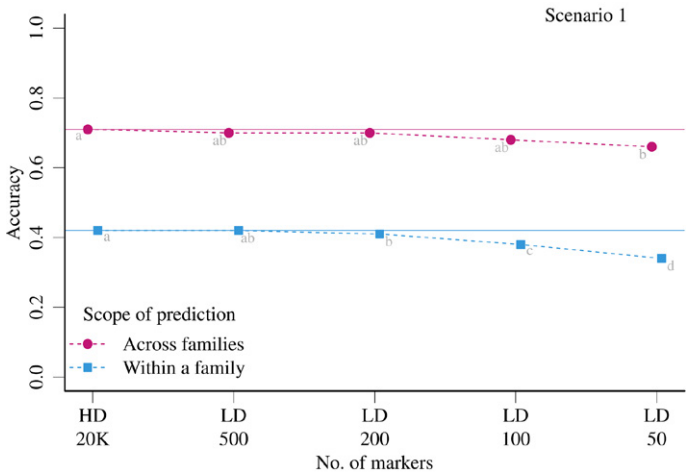

Fig. 2. Prediction accuracy across families and within a family against the number of markers used in the prediction set with 200 individuals; 2000 training individuals had the true high-density genotypes (HD); LD, low-density genotypes imputed to high density; letters denote significant difference within the scope of prediction at $p \leq 0.01$ according to the Tukey's multiple comparison test.

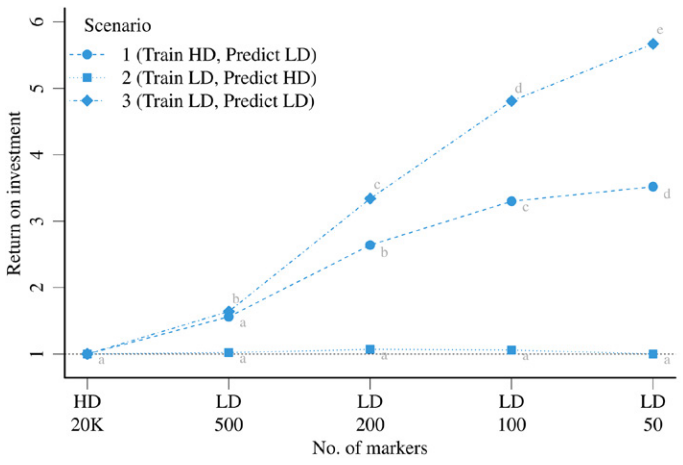

Fig. 3. Return on investment for selection within a family against the number of markers used in the training set with 2000 individuals and prediction set with 200 individuals; HD, true high-density genotypes; LD, low-density genotypes imputed to high density; letters denote significant difference within a scenario at $p \leq 0.01$ according to the Tukey's multiple comparison test.

Response to Selection with Imputation and Enlarged Prediction Set

Enlarging the prediction set through low-density genotyping and imputation increased response to selection through increased selection intensity. This is shown in Fig. 4, which plots the response to selection against the number of prediction individuals, i.e., the selection candidates. The selection candidates were evaluated based on the true high-density genotypes or low-density genotypes imputed to high density. Increasing the number of candidates increases response to selection. When the increased number of candidates was based on genotyping fewer markers, the response to selection stopped increasing after a certain number of low-density markers. Significant increase in response occurred when we increased the number of selection candidates from 50 genotyped at 500 low-density markers (response was 0.30) to 100 genotyped at 200 low-density markers (response was 0.36). Further increases in the number of selection candidates (above 100) while decreasing the number of low-density markers (below 200 markers) did not increase response any further.

Increasing response to selection through increased selection intensity was cost effective only with imputation. This is shown in Fig. 5, which plots the return on investment of selecting within a family against the number of prediction individuals, i.e., the selection candidates. The selection candidates were evaluated based on the true high-density genotypes or low-density genotypes imputed to high density. The baseline for comparison was the strategy from the first scenario, in which 200 selection candidates were genotyped at high density. When selection candidates had the true high-density genotypes, increasing selection intensity was not cost effective—the highest return on investment (2.05 times that of the baseline scenario) was achieved when 50 candidates were genotyped with 500 low-density markers. When selection candidates had imputed genotypes, increasing selection intensity was cost effective—the highest return on investment was achieved when 150 candidates were genotyped with 100 low-density markers, though this strategy was comparable with genotyping 100 candidates with 200 low-density markers or genotyping 200 candidates with 50 low-density markers. These three strategies gave return of investment between 3.39 and 3.68 times that of the baseline strategy.

Prediction Accuracy with Imputation and Enlarged Training Set

Enlarging the training set through low-density genotyping and imputation increased prediction accuracy. This is shown in Fig. 6, which plots the genomic prediction accuracy against the number of training individuals. The training individuals were from 80 families and had either the true high-density genotypes or low-density genotypes imputed to high density. Accuracies are shown both for prediction across families and within a family. Prediction accuracy increases with an enlarged training set. Predictions based on the imputed genotypes in training were of similar accuracy to those based on the true genotypes over a wide range of training set sizes and marker densities. For example, the highest loss of accuracy from 0.69 to 0.62 was observed when the scope of prediction was within a family and 50 low-density markers were used for imputation. The turning point at which imputed genotypes in training gave

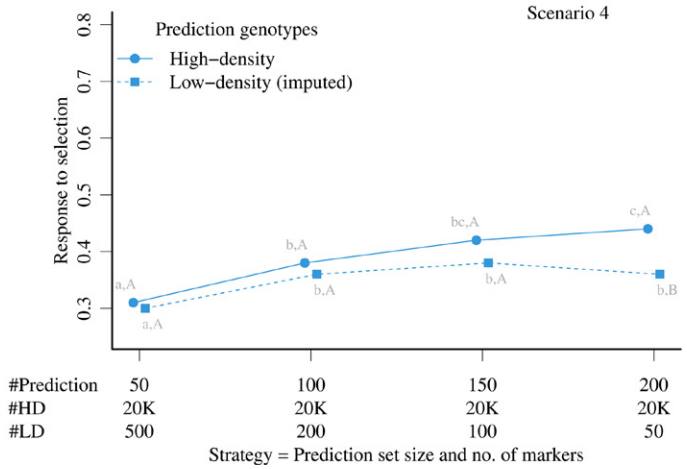

Fig. 4. Response to selection within a family against the number of selection candidates having the true high-density genotypes (HD) or low-density genotypes imputed to high density (LD; 2000 training individuals had the true high-density genotypes (points with a different letter—letters denote significant difference between strategies [the first letter] and used prediction genotypes [the second letter] at $p \leq 0.01$ according to the Tukey's multiple comparison test).

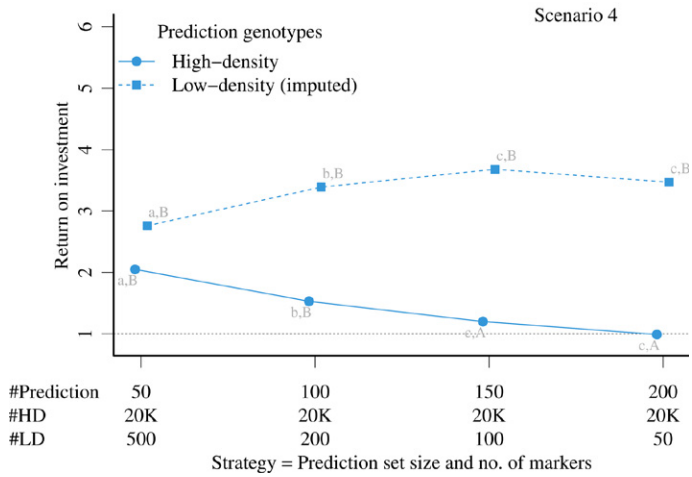

Fig. 5. Return on investment for selection within a family against the number of selection candidates having the true high-density genotypes (HD) or low-density genotypes imputed to high density (LD); 2000 training individuals had the true high-density genotypes (letters denote significant difference between strategies [the first letter] and the baseline scenario [the second letter] at $p \leq 0.01$ according to the Tukey's multiple comparison test).

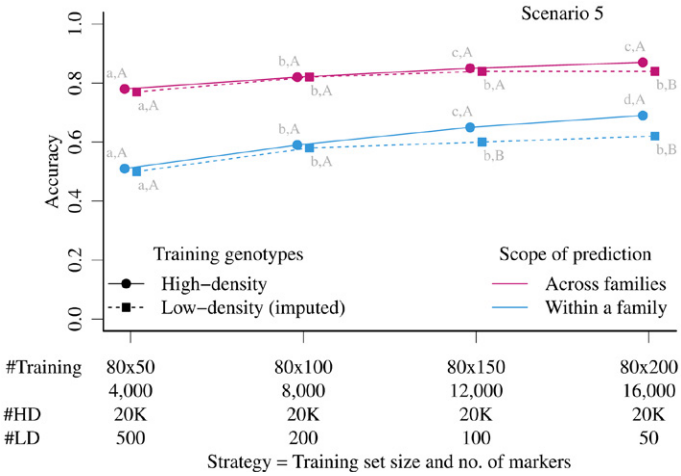

Fig. 6. Prediction accuracy across families and within a family against the number of training individuals from 80 families having the true high-density genotypes (HD) or low-density genotypes imputed to high density (LD); 200 prediction individuals had the true high-density genotypes (letters denote significant difference between strategies [the first letter] and used training genotypes [the second letter] at $p \leq 0.01$ according to the Tukey's multiple comparison test).

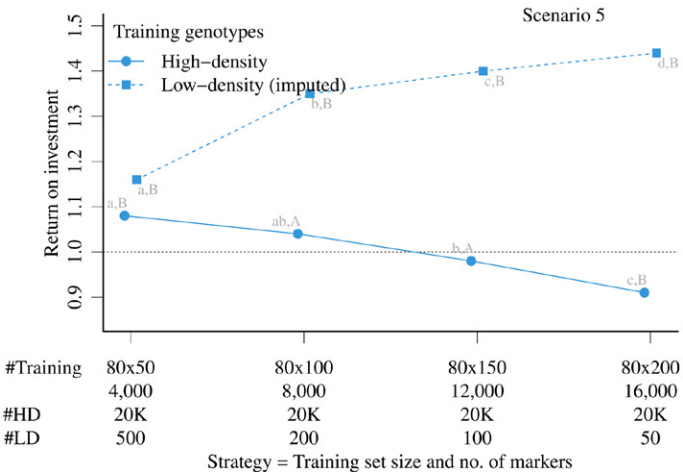

Fig. 7. Return on investment for selection within a family against the number of training individuals from 80 families having the true high-density genotypes (HD) or low-density genotypes imputed to high density (LD); 200 prediction individuals had the true high-density genotypes (letters denote significant difference between strategies [the first letter] and the baseline scenario [the second letter] at $p \leq 0.01$ according to the Tukey's multiple comparison test).

significantly lower prediction accuracy than the true genotypes differed between the scopes of prediction. When the scope of prediction was across families, the turning point was between 100 and 50 low-density markers. When the scope of prediction was within a family, the turning point was already between 200 and 100 low-density markers.

Enlarging the training set increased return on investment when low-density genotyping and imputation were used, but not when the true high-density genotypes were used. This is shown in Fig. 7, which plots the return on investment of selecting within a family against the number of training individuals. The training individuals had either the true high-density genotypes or low-density genotypes imputed to high density. The baseline for comparison was a strategy from the first scenario, in which 2000 training individuals were genotyped at high density. Doubling the baseline training set with high-density individuals gave 1.08 times higher return on investment, but further increases either gave comparable or lower return on investment than the baseline strategy. When low-density genotyping and imputation were used, the return on investment was larger and increased with increasing training set size. The maximal return on investment with that approach was 1.44 times that of the baseline strategy. This maximal scenario had a training size of 16,000 individuals genotyped at 50 low-density markers.

Response to Selection with Imputation and Enlarged Training and Prediction Sets

Enlarging both the training and prediction set through low-density genotyping and imputation increased response to selection with diminishing returns. This is shown in Fig. 8, which plots the response to selection within a family against the number of training and prediction individuals. Both sets of individuals had either the true high-density genotypes or low-density genotypes imputed to high density. Response to selection increased when both the number of training and prediction individuals with true high-density genotypes increased. However, when low-density genotypes and imputation were used, the increase in response to selection plateaued at strategies that used 100 low-density markers both in training with 12,000 individuals across 80 families and in prediction with 150 individuals per family.

Enlarging both the training and prediction set through low-density genotyping and imputation gave greater return on investment than using the true high-density

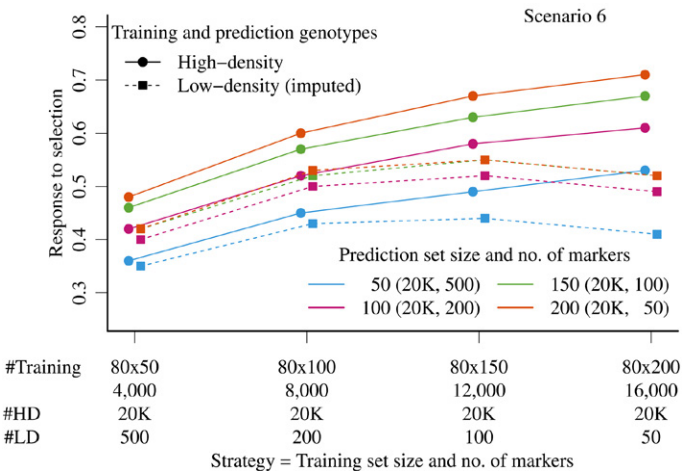

Fig. 8. Response to selection within a family against the number of training individuals from 80 families and prediction individuals having the true high-density (HD) genotypes or low-density (LD) genotypes imputed to high density.

genotypes. This is shown in Fig. 9, which plots the return on investment of selecting within a family against the number of training and prediction individuals. Both sets of individuals had either the true high-density genotypes or low-density genotypes imputed to high density. The baseline for comparison was a strategy from the first scenario, in which 2000 training individuals and 200 prediction individuals had high-density genotypes. When high-density genotypes were used, it was not cost effective to increase either the training set size or the prediction set. Using the imputation increased return on investment up to 5.12 times that of the baseline strategy. The most effective strategy was to assemble a training set of 12,000 individuals and a prediction set of 150 individuals per family, both genotyped at 100 low-density markers.

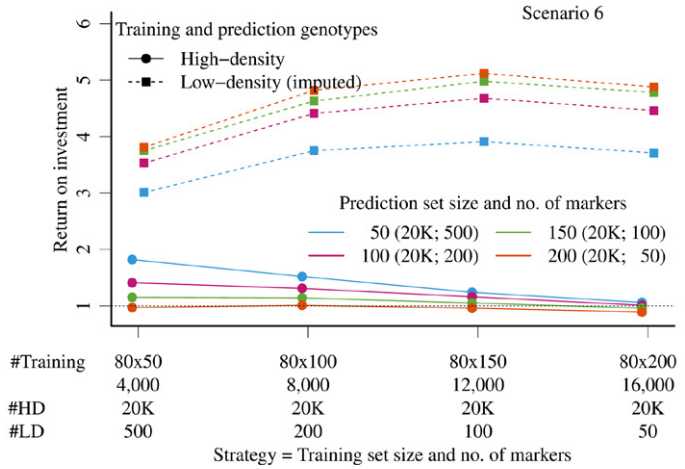

Fig. 9. Return on investment for selection within a family against the number of training individuals from 80 families and prediction individuals having the true high-density (HD) genotypes or low-density (LD) genotypes imputed to high density.

DISCUSSION

Our results highlight four main points for discussion, specifically (i) the three principles of cost effectively assembling the data for genomic selection through imputation, (ii) the required number of low-density genotypes, (iii) implications for breeding programs, and (iv) the assumptions made by the study.

The Three Principles of Cost Effectively Assembling the Data for Genomic Selection through Imputation

The results show that accurate within-family imputation can enable cost-effective genomic selection in plant breeding through three complementary principles: (i) reducing costs by low-density genotyping and imputation, (ii) increasing selection intensity by genotyping more candidates at fewer low-density markers, and (iii) increasing prediction accuracy by genotyping more training individuals at fewer low-density markers. Each of these principles is underpinned by phenomena that affect the power and cost effectiveness of a genomic selection program, and we discuss these in turn.

Reducing Genotyping Costs by Low-Density Genotyping and Imputation

Imputation is a technology designed to reduce the genotyping costs by exploiting the rules of inheritance on partially observed genotypes of relatives. In this study, we have leveraged this technology and the prevalent family structure of plant breeding programs to assemble the required data for genomic selection in a cost-effective way. Plant breeding programs are ideal for such an approach, because a strategy to genotype a small number of parents at high density and a large number of their progeny at low density enables large cost savings in genotyping the training or prediction sets for genomic selection.

When we used imputation only in the prediction set, the prediction accuracy was unimpaired when at least 200 low-density markers were used. However, the return on investment was optimal with 50 low-density markers (3.52 that of the baseline scenario), because at that density, the 19% loss in prediction accuracy (= 1 − 0.34/0.42, Supplemental Table S2) was more than compensated by the 87% cost reduction in genotyping an individual (= 1 − $4.00/$30.00, Supplemental Table S2). When we used imputation only in the training set, the prediction accuracy was reduced less, but the return on investment did not improve compared with the baseline scenario. This was because we spread the cost of genotyping the training set across a large number of predictions and the reduction in that cost was negligible when analyzed on a per-family basis. The return on investment was the greatest when imputation was used both in training and prediction sets, because this scenario enabled the greatest total cost reduction in genotyping. The most optimal strategy was to genotype the sets with 50 low-density markers, which gave 5.67 times more return on investment than the baseline scenario. This was achieved through a 24% loss in prediction accuracy (= 1 − 0.34/0.45, Supplemental Table S2) but an 87% cost reduction in genotyping the sets (= 1 − $2400/$18,000, Supplemental Table S2).

Of note is the observation that, when imputed genotypes were used only in the training set, the prediction accuracy decreased at a lower rate than when imputed genotypes were used only in the prediction set. This can be explained by the fact that, for training a genomic selection equation, a set of individuals is used to estimate population parameters (variance components and allele substitution effects) and errors for any individual average out across the set of individuals to some degree. On the other hand, in the prediction set, each individual is evaluated independently and is more dependent on having few errors in its genotype (e.g., Gorjanc et al., 2015). This observation suggests that a

small amount of errors introduced by imputation has larger effect on prediction than on training.

Increasing Selection Intensity by Genotyping More Candidates at Fewer Low-Density Markers

Increasing selection intensity increases response to selection, but the benefit must be balanced against increased costs of evaluating more selection candidates. Our results show that this is hard to achieve with high-density genotyping but is possible with low-density genotyping and imputation. For example, increasing selection intensity from 1.4 (10 selected candidates out of 50 candidates) to 2.1 (10 selected candidates out of 200 candidates) increased response to selection by 33% when the candidates were genotyped at high density (= 1 − 0.31/0.44, Supplemental Table S2) and by 17% when the candidates were genotyped at low density and imputed (= 1 − 0.30/0.36, Supplemental Table S2). However, this strategy increased the return on investment only when low-density genotyping and imputation were used, because the total costs of genotyping an increasing number of selection candidates at fewer markers was nearly constant (between $800 and $875, Supplemental Table S2). The return on investment did not increase when the true high-density genotypes were used, because the total costs of genotyping increased linearly with an increasing number of selection candidates (from $1500 to 6000, Supplemental Table S2) and outweighed the benefit of an increased response to selection.

The observed dynamic can be explained by the fact that, while selection intensity increases at an increasing (nonlinear) rate against a decreasing proportion of selected individuals, it increases almost linearly for a wide range of proportions, i.e., when >20% of candidates are selected (Falconer and Mackay, 1996). This means that a greater response to selection through more intense selection must outweigh greater costs of evaluating more selection candidates at a rate that is more than linear. This can only be achieved with genotyping an increasing number of candidates at ever fewer markers. However, this strategy can only be used to the point where the loss in prediction accuracy due to imputation errors diminishes response to selection and the higher cost of genotyping ever more individuals outweighs the benefit. In this study, this point was observed when 10 candidates were selected out of 150 that were genotyped at 100 low-density markers. This scenario gave the return on investment of 3.68 times that of the baseline scenario (genotyping the 200 candidates at high density).

Increasing Prediction Accuracy by Genotyping More Training Individuals at Fewer Low-Density Markers

Increasing accuracy of selection through enlarging the training set increases response to selection, but this principle must also be balanced with the costs of achieving that level of accuracy. Our results show that low-density genotyping and imputation enable increasing response to selection with this strategy in a cost-effective way, while high-density genotyping leads to overinvestment in genotype data. Since accuracy of imputation was high in this study, it is not surprising that imputation enabled a cost-effective way to increase the size of the training set and with that the prediction accuracy. However, the overinvestment with high-density genotyping was surprising. There are at least two phenomena that underlie this observation. First, while the genomic prediction accuracy increases with an increasing training set size, it does so with diminishing returns (Daetwyler et al., 2008; Goddard, 2009). Since our baseline training set of 2000 individuals was already sizeable, it is expected that increases in the training set do not increase prediction accuracy substantially. When this is coupled with the higher costs of genotyping more training individuals at high density, it leads to an expectation that the return on investment reduces. Second, the design of our simulation likely exacerbated the first phenomena. Namely, we have increased the baseline training set by sampling more individuals from the same families. In addition, each family in the prediction set had, on average, 7.7 families with one parent in common in the training set. These two design properties imply that the baseline training set already covered most of the genetic variability and that the relationship between the training and prediction sets was high (Clark et al., 2012; Pszczola et al., 2012; Hickey et al., 2014). Such high coverage of genetic variability and high connectedness between the training and prediction sets is not expected for every breeding program, especially when rapid cycling is aggressively used. In such cases, increasing or updating the training set every year is essential to maintain prediction accuracy (Michel et al., 2016; Pszczola and Calus, 2016). Our work shows that this can be achieved in a cost-effective way with low-density genotyping and imputation.

Required Number of Low-Density Markers

The required number of low-density markers for cost-effective genomic selection depends on many parameters and a setting where the data will be used. The results of this study suggest that imputation with about 20 segregating markers per chromosome gives comparable prediction accuracy and response to selection as high-density genotypes. However, in the terms of return on investment, the number can be reduced to as few as five segregating markers per chromosome when imputation is used both in the training and prediction sets with or without enlarging the sets. These numbers refer to the number

of segregating markers, and unless marker platforms can be cost effectively developed for each family specifically, a greater number of assayed markers will be needed to ensure so many segregating markers in a family. Assuming that about one third of markers segregate in a family (e.g., Hickey et al., 2014), the targeted number of low-density markers should be between 15 and 60 per 1-Morgan chromosome or between 150 and 600 per genome with 10 1-Morgan chromosomes. That so few markers are sufficient is consistent with previous results from studies in simulated (Hickey et al., 2015) and real data (Jacobson et al., 2015). Hickey et al. (2015) discuss in detail why plant breeding populations enable accurate imputation with so few low-density markers. Here, we emphasize that a strategy of densely genotyping a small number of parents and sparsely genotyping a large number of their progeny enables large reductions in the total genotyping costs, while response to selection is not diminished substantially.

Implications for Breeding Programs

Our study has three implications for plant-breeding programs:

First, the results show that large cost reductions can be achieved with the proposed genotyping strategy and accurate within-family imputation. This is extremely important, because the cost of assembling the required data is the key-limiting factor for adopting genomic selection. We have shown that this limitation can be overcome by (i) lowering the costs of assembling sufficiently large training sets that yield accurate predictions in unphenotyped families and (ii) lowering the cost of assembling large prediction sets that yield measurable response to selection.

Second, low-cost genotyping enables adoption of genomic selection in early segregating populations. In that stage of the breeding program, the potential of genomic selection is likely to be the greatest, because breeders could select early for all traits covered by the training set, even yield. However, that stage is also the most challenging for implementing genomic selection, because early segregating populations comprise large number of individuals. Genotyping costs should be as low as possible to make this a possibility. Genomic prediction at that stage could be combined with a prior phenotype screening for traits that are inexpensive to measure. This strategy would avoid the need to genotype individuals with poor phenotypes for these traits. Our results show that genotyping with 50 low-density markers and imputing can give accuracy of prediction within an unphenotyped family of at least 0.3. Coupling this level of accuracy with large genetic variance in segregating populations and short generation interval promises substantial responses to selection, which could be achieved in a cost-effective way.

Third, low-density genotyping and imputation affect accuracy of prediction in a different way when the scope of prediction is across families or within a family. Our results show that prediction accuracy within a family is more sensitive to imputation errors than prediction accuracy across families. This is expected, because prediction accuracy across families is due to capturing the between- and within-family genetic variation, while prediction accuracy within a family is only due to capturing the within-family genetic variation. It is easy to accurately impute the part of genotypes that is due to between-family variation, i.e., the mean genotype of the parents. It is much more challenging to accurately impute the part of genotypes that is due to within-family variation, i.e., the deviation of progeny's genotype from the mean genotype of the parents. This is important, because imputation accuracy and the resulting prediction accuracy with imputed genotypes are influenced by the population or family structure in the same way as prediction accuracy with non-imputed genotypes is (e.g., Windhausen et al., 2012). Ignoring this phenomenon can lead to a breeding program that underuses the potential of genomic selection to capture within-family variation. This is important also in the terms of long-term gain and sustainability of a breeding program, which depend largely on the ability to select on the within-family variation, while selection on the between-family variation leads to rapid depletion of genetic variation (e.g., Woolliams et al., 1999, 2015).

Assumptions of the Study

The estimated benefit of imputation for cost-effective genomic selection in plant breeding depends on some assumptions made in this study. Breeding programs attempting to follow the described approach might want to reevaluate the benefit of imputation by varying the population structure, the size of a program, and most importantly the costs. We have modelled the cost of genotyping based on inquiries from several genotyping providers, in particular from LGC (http://www.lgcgroup.com). The assumed costs were for a species with an established genome sequence and a fairly large breeding program that would enable the economy of scale. We have spread the cost of assembling the training genotypes over one cycle of predictions in 80 families, while the cost of the prediction genotypes was attributed to each family. We believe this is a conservative approach, because we do not expect large genetic changes so that the training set would have to be fully replaced after one cycle of selection. This suggests that using imputation in training might be even more cost effective than suggested in this study. If costs are spread over many cycles of selection, drop in prediction accuracy with distancing generations should be accounted for. More critically, we have assumed that phenotype data is available and that a starting genomic selection program could simply reuse this data. We believe this is a reasonable assumption for existing breeding programs, but if this assumption is not met, the return on investment can change considerably. The spreadsheet provided in supplement can be used to change our assumptions and corroborate our results under different settings.

We have performed sensitivity analysis by doubling low-density genotyping costs (both fixed and variable parts or just one of the two) and found that, while the values for the return on investment change, the relative comparison of the evaluated strategies in the scenarios does not change. Specifically, doubling the fixed and variable costs of low-density genotyping increased total costs of 50 markers from \$4.00 to 8.00, of 100 markers from \$5.50 to 11.00, and of 200 markers from \$8.50 to 17.00. In scenarios 1–3 (fixed size of training and prediction sets), these cost changes have reduced the return on investment for the optimal scenario (low-density genotyping with 50 markers and imputation both in training and prediction sets) from 5.67 to 2.84 (a 50% reduction), but this scenario was still the most optimal. In scenario 4 (enlarging the prediction set by low-density genotyping and imputation), these cost changes reduced the return on investment for 35% in all settings. In scenario 5 (enlarging the training set by low-density genotyping and imputation), these cost changes reduced the return on investment for 11% in all settings. In scenario 6 (enlarging both training and prediction sets by low-density genotyping and imputation), these cost changes reduced the return on investment for 50% in all settings.

The reviewers pointed out that our simulation design does not resemble an evolving breeding program and that <20,000 high-density markers might be sufficient for the same prediction accuracy but have a lower cost, and hence a greater return on investment. The simulation design used is not an evolving breeding program, but it is indicative of a snapshot of such a program at one time point. The founder chromosomes were sampled from a coalescent process with effective population size of 50 (with increasing values in the past). Therefore, there was a trajectory of relationships between the founders and families that is representative of what may be present in a particular breeding program, and our results show average over this trajectory. Hence, our results inform about the potential of imputation to lower the cost of genotyping large number of individuals for genomic selection, but each breeding program should evaluate this potential for its specific conditions.

We agree that <20,000 high-density markers are required for accurate predictions among closely related individuals. While parents could have been genotyped with fewer high-density markers, this might not necessarily reduce costs considerably, as there is a nonlinear relationship between the number of markers and the cost, in particular when progressing between the low-density and high-density types of arrays. It should be emphasized that the cost of high-density genotypes on a relatively small number of parents is only a fraction of the total required genotyping budget for genomic selection, hence not primary target for cost optimization. Also, using a surplus of high-density markers is beneficial for at least two reasons. First, when the relationship between the training and prediction sets reduces, the required number of markers to achieve a targeted level of accuracy increases (e.g., Hickey et al., 2014). By having high-density markers, we ensure that, in the longer term, the training set can be used for making selection decisions in more families and thus have its cost of construction offset over more selection decisions. Second, surplus of high-density markers ensures good coverage of germplasm genetic diversity and reduces ascertainment bias (e.g., Ganal et al., 2012; Heslot et al., 2013).

CONCLUSION

Accurate within-family imputation enables cost-effective genomic selection in plant breeding. This can be achieved through (i) reduced cost of genotyping the training and prediction sets by low-density genotyping and imputation, (ii) increased selection intensity by genotyping more selection candidates at fewer markers and imputing, and (iii) increased genomic prediction accuracy by genotyping more training individuals at fewer markers and imputing. These three principles enable plant breeders to cost effectively assemble the required data for genomic selection.

Conflict of Interest

The authors declare there to be no conflict of interest.

Supplemental Material Available

Supplemental material for this article is available online.

Acknowledgments

The authors acknowledge the financial support from the BBSRC ISPG to The Roslin Institute BB/J004235/1, from Genus PLC and from grant numbers BB/M009254/1, BB/L020726/1, BB/N004736/1, BB/N004728/1, BB/L020467/1, and BB/N006178/1 and Medical Research Council (MRC) grant number MR/M000370/1. The authors thank Dr. Andrew Derrington (Scotland, UK) for assistance in refining the manuscript. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/).

References

Bernardo, R., and J. Yu. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 47:1082–1090. doi:10.2135/cropsci2006.11.0690

Calus, M.P.L., A.C. Bouwman, J.M. Hickey, R.F. Veerkamp, and H.A. Mulder. 2014. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: A review of livestock applications. Animal 8:1743–1753. doi:10.1017/S1751731114001803

Chen, G.K., P. Marjoram, and J.D. Wall. 2009. Fast and flexible simulation of DNA sequence data. Genome Res. 19:136–142. doi:10.1101/gr.083634.108

Clark, S.A., J.M. Hickey, H.D. Daetwyler, and J.H. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Sel. Evol. 44(1):4. doi:10.1186/1297-9686-44-4

Cleveland, M.A., and J.M. Hickey. 2013. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. J. Anim. Sci. 91:3583–3592. doi:10.2527/jas.2013-6270

Daetwyler, H.D., B. Villanueva, and J.A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One 3(10):e3395. doi:10.1371/journal.pone.0003395

Endelman, J.B., G.N. Atlin, Y. Beyene, K. Semagn, X. Zhang, M.E. Sorrells, and J.-L. Jannink. 2014. Optimal design of preliminary yield trials with genome-wide markers. Crop Sci. 54:48–59. doi:10.2135/cropsci2013.03.0154

Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to quantitative genetics. Longman, Harlow, UK.

Faux, A.-M., G. Gorjanc, R.C. Gaynor, M. Battagin, S.M. Edwards, D.L. Wilson et al. 2016. AlphaSim: Software for breeding program simulation. Plant Genome 9(3):1–14. doi:10.3835/plantgenome2016.02.0013

Ganal, M.W., A. Polley, E.-M. Graner, J. Plieske, R. Wieseke, H. Luerssen, and G. Durstewitz. 2012. Large SNP arrays for genotyping in crop plants. J. Biosci. 37:821–828. doi:10.1007/s12038-012-9225-3

Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica (The Hague) 136:245–257.

Gorjanc, G., M.A. Cleveland, R.D. Houston, and J.M. Hickey. 2015. Potential of genotyping-by-sequencing for genomic selection in livestock populations. Genet. Sel. Evol. 47:12. doi:10.1186/s12711-015-0102-z

Heffner, E.L., J.-L. Jannink, and M.E. Sorrells. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome 4:65–75. doi:10.3835/plantgenome.2010.12.0029

Heslot, N., J. Rutkoski, J. Poland, J.-L. Jannink, and M.E. Sorrells. 2013. Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. PLoS One 8(9):e74612. doi:10.1371/journal.pone.0074612

He, S., Y. Zhao, M.F. Mette, R. Bothe, E. Ebmeyer, T.F. Sharbel et al. 2015. Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat (*Triticum aestivum* L.). BMC Genomics 16(1):168. doi:10.1186/s12864-015-1366-y

Hickey, J.M., J. Crossa, R. Babu, and G. de los Campos. 2012a. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52:654–663. doi:10.2135/cropsci2011.07.0358

Hickey, J.M., S. Dreisigacker, J. Crossa, S. Hearne, R. Babu, B.M. Prasanna et al. 2014. Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci. 54:1476–1488. doi:10.2135/cropsci2013.03.0195

Hickey, J.M., G. Gorjanc, R.K. Varshney, and C. Nettelblad. 2015. Imputation of single nucleotide polymorphism genotypes in biparental, backcross, and topcross populations with a hidden Markov model. Crop Sci. 55:1934–1946. doi:10.2135/cropsci2014.09.0648

Hickey, J.M., B.P. Kinghorn, B. Tier, J.H. van der Werf, and M.A. Cleveland. 2012b. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet. Sel. Evol. 44(9):11.

Hoerl, A.E., and R.W. Kennard. 1976. Ridge regression iterative estimation of the biasing parameter. Commun. Stat. Theory Methods 5:77–88. doi:10.1080/03610927608827333

Huang, Y., J.M. Hickey, M.A. Cleveland, and C. Maltecca. 2012. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. Genet. Sel. Evol. 44(1):25. doi:10.1186/1297-9686-44-25

Jacobson, A., L. Lian, S. Zhong, and R. Bernardo. 2014. General combining ability model for genomewide selection in a biparental cross. Crop Sci. 54:895–905. doi:10.2135/cropsci2013.11.0774

Jacobson, A., L. Lian, S. Zhong, and R. Bernardo. 2015. Marker imputation before genomewide selection in biparental maize populations. Plant Genome 8(2):1–9. doi:10.3835/plantgenome2014.10.0078

Lian, L., A. Jacobson, S. Zhong, and R. Bernardo. 2014. Genome-wide prediction accuracy within 969 maize biparental populations. Crop Sci. 54:1514–1522. doi:10.2135/cropsci2013.12.0856

Li, Y., C.J. Willer, J. Ding, P. Scheet, and G.R. Abecasis. 2010. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34:816–834. doi:10.1002/gepi.20533

Lorenzana, R.E., and R. Bernardo. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. 120:151–161. doi:10.1007/s00122-009-1166-3

Mackay, I.J., P. Bansept-Basler, T. Barber, A.R. Bentley, J. Cockram, N. Gosman et al. 2014. An eight-parent multiparent advanced generation inter-cross population for winter-sown wheat: Creation, properties, and validation. G3: Genes, Genomes, Genet. 4:1603–1610. doi:10.1534/g3.114.012963

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Michel, S., C. Ametz, H. Gungor, D. Epure, H. Grausgruber, F. Löschenberger, and H. Buerstmayr. 2016. Genomic selection across multiple breeding cycles in applied bread wheat breeding. Theor. Appl. Genet. 129:1179–1189. doi:10.1007/s00122-016-2694-2

Nicholas, F.W. 1980. Size of population required for artificial selection. Genet. Res. 35:85–105. doi:10.1017/S0016672300013951

Pszczola, M., and M.P.L. Calus. 2016. Updating the reference population to achieve constant genomic prediction reliability across generations. Animal 10:1018–1024. doi:10.1017/S1751731115002785

Pszczola, M., T. Strabel, H.A. Mulder, and M.P.L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95:389–400. doi:10.3168/jds.2011-4338

R Development Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rutkoski, J.E., J. Poland, J.-L. Jannink, and M.E. Sorrells. 2013. Imputation of unordered markers and the impact on genomic selection accuracy. G3: Genes, Genomes, Genet. 3:427–439. doi:10.1534/g3.112.005363

Schaeffer, L.R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218–223. doi:10.1111/j.1439-0388.2006.00595.x

Whittaker, J.C., R. Thompson, and M.C. Denham. 2000. Marker-assisted selection using ridge regression. Genet. Res. 75:249–252. doi:10.1017/S0016672399004462

Windhausen, V.S., G.N. Atlin, J.M. Hickey, J. Crossa, J.-L. Jannink, M.E. Sorrells et al. 2012. Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3: Genes, Genomes, Genet. 2:1427–1436. doi:10.1534/g3.112.003699

Woolliams, J.A., P. Berg, B.S. Dagnachew, and T.H.E. Meuwissen. 2015. Genetic contributions and their optimization. J. Anim. Breed. Genet. 132:89–99. doi:10.1111/jbg.12148

Woolliams, J.A., P. Bijma, and B. Villanueva. 1999. Expected genetic contributions and their impact on gene flow and genetic gain. Genetics 153:1009–1020.

Xavier, A., W.M. Muir, and K.M. Rainey. 2016. Impact of imputation methods on the amount of genetic variation captured by a single-nucleotide polymorphism panel in soybeans. BMC Bioinformatics 17:55. doi:10.1186/s12859-016-0899-7